Cluster analysis

Uses of cluster analysis

- Cluster analysis solves the following problem:
  - Given a sample of n objects, each with a score on p variables, devise a scheme for grouping the objects into classes so that '*similar*' ones are in the same class.
  - The method must be completely numerical and the number of classes is not known.
- Reasons why cluster analysis may be worthwhile.
  - For finding the 'true' groups, or, it may be useful for data reduction. For example, a large number of cities can potentially be used as test markets for a new product but it is only feasible to use a few. If cities can be grouped into a small number of groups of similar cities then one member from each group could be used for the test market.

Types of cluster analysis

- Many algorithms have been proposed for cluster analysis.
- The *hierarchic techniques* produce a dendrogram such as the ones shown in the family size example. These methods start with the calculation of the distances of each individual to all other individuals. With agglomeration all objects start by being alone in groups of one. Close groups are then gradually merged until finally all individuals are in a single group.
- In the *K-means* method the algorithm assigns each item to the cluster having the nearest centroid (mean).